Invention Journal of Research Technology in Engineering & Management (IJRTEM)
ISSN: 2455-3689
www.ijrtem.com Volume 2 Issue 3 ǁ March. 2018 ǁ PP 01-06

E-Symptom Analysis System to Improve Medical Diagnosis and Treatment Recommendation

Merve Kevser Gökgöl [1] , Zeynep Orhan [2]
[1,2]*Department of Information Technologies, International Burch University, Sarajevo, Bosnia and Herzegovina*

ABSTRACT: *A wealth of data in public health care systems has been collected and meanwhile there are plenty of new technological improvements which have considerable influence on current data pool. Nevertheless, important obstacles are challenging to utilize existing clinical data. Enhanced technological improvements lead patients to search their symptoms and corresponding diagnosis on online resources. In this study, it is aimed to develop a machine learning model to suit in different availability of users. Most of the current systems allow people to choose related symptom in web interfaces or Q&A forums. In addition to these applications it is aimed to implement a new technique which extracts the text-based symptoms and its related parameters such as, severity, duration, location, cause, accompanied by any other indicators. This study is applicable for patient`s everyday language statements besides medical expression of symptoms for corresponding symptoms. Extracted terms are used as an input of the model and analyzed for matching diagnosis where an accuracy of 72.5% has been accomplished.*

KEYWORDS: *Symptom extraction, text matching, medical diagnosis, public healthcare*

I. INTRODUCTION

The era of digitization has led computers to become the real face of handling commercial processes across a wealth of industries. As institutions today specifically in the medical domain have resorted to these virtual machines for realizing their goals, more and more medical data is being generated on a continuous basis (Davis et al., 2008). This data is being used in many recommendation systems which deliver a personalized individual's health profile model (Bolón-Canedo et al., 2015). The quality of health care can be precisely defined and measured with a degree of scientific accuracy comparable with that of most measures used in clinical medicine (Schiff, 2009). Every year very large number of population is affected by wrong or late diagnosis (Leape et al., 1991). Diagnoses that are missed, incorrect or delayed are believed to affect 10 to 20 percent of cases, far exceeding drug errors and surgery on the wrong patient or body part, both of which have received considerably more attention (Leape et al., 1991).

Current systems are challenging to optimize the utilization of existing medical data resources in automated diagnostics. With an increased significance on providing better quality and reducing costs, new systems are required to improve in public healthcare. Clinical decision support systems (CDSS) are computer systems designed to impact clinical decision making about individual patients at the point in time that these decisions are made (Kaul, Kaul, & Verma, 2015). It is also important to balance prevention of medical errors while providing a quick and low-cost health services. CDSS have been a key element of systems' approaches to improving patient safety and the quality of care and have been a key requirement for "meaningful use" of electronic health records (EHRs) (Kaul et al., 2015). The potential for information technology in health care is still in the process of being actualized. Large dimensionality of data in medicine together with the common reduced sample size of pathological cases makes indispensable the use of advanced machine learning techniques for clinical interpretation and analysis (Spyns, Nhàn, Baert, Sager, & De Moor, 1998). The detection and interpretation of pathological conditions usually required large number of experts available, however the number of experts is sometimes not enough, and other problems may appear such as disagreement among experts (Spyns et al., 1998).

Information technology (IT) maintains a significant, sustainable knowledge which is vital for organization development. While the utilization of current data leads services to be more accessible and mobile, it is vital for health care industry to provide easier and faster accessibility as well as affordable and higher quality of service. Health IT is not just about merely digitizing medical records to create a paperless office, although doing this will achieve considerable savings, it is also about fundamentally transforming the health care system so that both doctors and patients have access to information and tools that allow them to better manage their care (Singh & Sittig, 2015).

II. BACKGROUND

There are a wide range of techniques that can be applied to analyzing these texts, as reflected in the considerable amount of research in the field of natural language processing (Popowich, 2005) The electronic patient records contains a rich source of valuable clinical information, which could be used for a wide range of automated applications aimed at improving the health care process, such as alerting for potential medical errors, generating a patient problem list, and assessing the severity of a condition (Friedman et al., 2004). However, these applications are not applicable since large amount of information is in textual form (Tange et al., 1998). Techniques for automatically encoding textual documents from the medical record have been evaluated by several groups. Examples are the Linguistic String Project (Xu et al., 2010), and Medical Language Extraction and Encoding system (Med LEE) (Friedman et al., 2004). Med LEE has been recently adapted to extract Unified Medical Language System (UMLS) concepts from medical text documents, achieving 83% recall and 89% precision (Alan R Aronson, 2001).

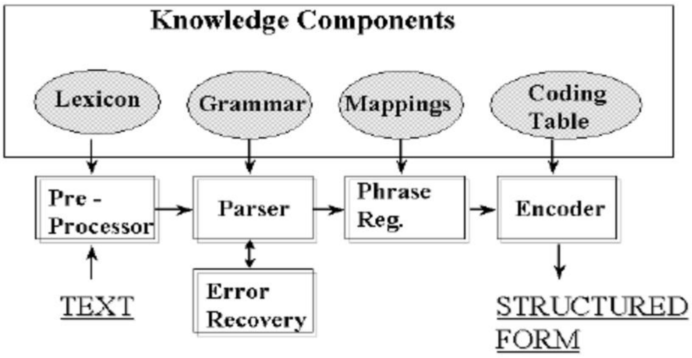

Figure 1. Overview of components of Med LEE

Figure 1 indicates the components of Med LEE. The knowledge-based components are shown as ovals; the processing engines are shown as rectangles and the new work discussed in this report involves the final stages of processing, the encoding process, which occurs after structured output is obtained (Friedman et al., 2004) Other systems automatically mapping clinical text concepts to a standardized vocabulary have been reported, like Meta Map (Zou et al., 2003), Index Finder (Pratt et al., 2003), and Knowledge Map (A R Aronson & Rindflesch, 1997). Meta Map and its Java version called Meta Map Transfer (MM Tx) were developed by the US National Library of Medicine (NLM). They are used to index text or to map concepts in the analyzed text with UMLS concepts.

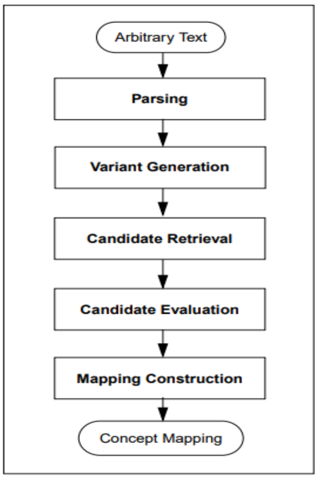

Figure 2. Meta Map sample of processing

A sample Meta Map processing biomedical text and to extract different types of information like anatomical concepts or molecular binding concepts (Wright et al., 1999) is shown in Figure 2. Meta Map has also been used with patient's electronic messages to automatically provide relevant health information to the patients (Meystre et al., 2008).

III. METHODOLOGY

In this study, process of data collection allows patients to enter their symptoms by typing in everyday language. Therefore, to increase the accuracy firstly it is required to clean and eliminate significant words. During this preprocessing, stop words, vague abbreviations are removed. Each symptom and other data valued words including severity, duration, location, cause, accompanied by any other symptoms, change in intensity are also extracted from written expression (as an individual expression or sentence structure of symptoms) accordingly. The structure of collected data categorized in four main branches; symptoms, diseases, tests (medical examinations) and corresponding treatments as described in Table 1.

Symptoms	Diseases	Tests	Treatments
ID	ID	ID	ID
Name	Name	Name	Name
No cation ID	Description	Definition	Definition
Level ID	Symptoms ID	Why is done	
Causes	Test ID	Preparation	
See doctor	Treatment ID	Expectation	
	Risks	Risks	
	Causes	Results	
	Prevention		
	Complications		
	Gender		
	Age		

Table 1. General structure of database table

Once tables are created symptoms are analyzed as input and they are trained by the process to detect possible diagnoses, tests and treatments. Disease data table as indicated in Table 2 consists of ID, Name, Description, Symptoms ID, Tests ID, Treatments ID, Risks, Causes, Preventions and Complications. Data content of symptoms, tests and treatments are represented with numerical values in different databases and they are embedded to system to analyze strength of the relationship.

ID	24
Name	Leukemia
Description	Leukemia is cancer of the body\'s blood-forming tissues, including the bone marrow and the lymphatic system.
Symptoms ID	21,69,114,119,133,168,173,0,0,0,0,0,0,0,0
Test ID	8,94,104
Treatment ID	113,114,115,116,117
Risks	Factors that may increase your risk of developing some types of leukemia include: Previous cancer treatment. People who\'ve had certain types of chemotherapy and radiation therapy for other cancers have an increased risk of developing certain types of leukemia. Genetic disorders. Genetic abnormalities seem to play a role in the development of leukemia. Certain genetic disorders, such as Down syndrome, are associated with an increased risk of leukemia. Exposure to certain chemicals. Exposure to certain chemicals, such as benzene â€” which is found in gasoline and is used by the chemical industry â€” also is linked to an increased risk of some kinds of leukemia. Smoking. Smoking cigarettes increases the risk of acute myelogenous leukemia. Family history of leukemia. If members of your family have been diagnosed with leukemia, your risk for the disease may be increased'
Causes	Scientists don\'t understand the exact causes of leukemia. It seems to develop from a combination of genetic and environmental factors.
Prevention	None
Complications	None
Gender	Not gender restricted
Age	Not age restricted

Table 2. Diseases table in database

As shown in Table 3 symptoms are identified by ID value, name, locations (neck, knee, eye, low back), level of the symptom which is also indicated with numerical values (sharp, low, high, sudden, mild, etc…) and causes, and possible conditions for a patient to visit the doctor.

ID	3
Name	Abdominal pain
Location ID	stomachache, tummy ache, gut ache and bellyache
Level ID	49, 9
Causes	Abdominal pain that steadily worsens over time, often accompanied by the development of other symptoms, is usually serious. Causes of progressive abdominal pain include: Cancer, Crohn\'s disease, Enlarged spleen (splenomegaly), Gallbladder cancer, Hepatitis (liver inflammation), Kidney cancer, Lead poisoning, Liver cancer, Non-Hodgkin\'s lymphoma, Pancreatic cancer, Stomach cancer, Tubo-ovarian abscess (pus-filled pocket involving a fallopian tube and an ovary), Uremia (buildup of waste products in your blood).
See doctor	Have someone drive you to urgent care or the emergency room if you have: Severe pain, Fever, Bloody stools, Persistent nausea and vomiting, Weight loss, Skin that appears yellow, Severe tenderness when you touch your abdomen, Swelling of the abdomen.

Table 3. A sample table of symptoms

Table of tests (medical examinations) which is given in Table 4, describes details about essential examinations` ID, name, definition, why it is required (why is done), preparation, expectation, risks, results.

ID	5
Name	Bilirubin test
Definition	Bilirubin testing checks for levels of bilirubin in your blood.
Why is done	Bilirubin testing is usually done as part of a group of tests to check the health of your liver. Bilirubin testing may be done to: Investigate jaundice â€" elevated levels of bilirubin can cause yellowing of your skin and the whites of your eyes (jaundice). A common use of the test is to measure bilirubin levels in newborns, determine whether there might be blockage in your liver\'s bile ducts, help detect or monitor the progression of other liver disease, such as hepatitis, help detect increased destruction of red blood cells, help follow how a treatment is working, help evaluate suspected drug toxicity.
Preparation	None
Expectation	Bilirubin testing is done using a blood sample. Usually, the blood is drawn through a small needle inserted into a vein in the bend of your arm. The needle is attached to a small tube, in which your blood is collected.You may feel a quick pain as the needle is inserted into your arm and experience some short-term discomfort at the site after the needle is removed.
Risks	None
Results	Normal results for a bilirubin test are 1.2 milligrams per deciliter (mg/dL) of total bilirubin for adults, and usually 1 mg/dL for those under 18. Normal results for direct bilirubin are generally 0.3 mg/dL.')

Table 4. Tests table in database

Recommended treatments for the most related symptoms are demonstrated as in Table 5. Treatments are classified with as their ID numbers, names and definitions.

ID	1
Name	Open abdominal surgery'
Definition	If you have an abdominal aortic aneurysm, surgery is generally recommended if your aneurysm is about 1.9 to 2.2 inches (about 5 to 5.5 centimeters) or larger. Open abdominal surgery to repair an abdominal aortic aneurysm involves removing the damaged section of the aorta and replacing it with a synthetic tube (graft), which is sewn into place. This procedure requires open abdominal surgery, and it will generally take you a month or more to fully recover.

Table 6. Table of treatments in database

Implementations: An Intelligent context utilizing recommendation engine (I CURE) for medical diagnosis is developed to convert the clinical data into significant and effective information. Python is used to develop the most efficient and appropriate model. Text Blob is a library used for input and output processing, and for string matching which actually classifies input symptoms as a disease. Two sets are used for classification, one including only symptoms, and the other the matching diseases. Details about diseases such as treatments and tests are recommended, are stored in MySQL database. Firstly, the user is asked to enter his/her symptoms, and then the application converts the answer into Text Blob object. Fuzzy String Matching, also called approximate string matching, is implemented as the process of finding strings which approximately match a given pattern. The closeness of a match is often measured in terms of edit distance which is the number of primitive operations necessary to convert the string into an exact match.

Training: In training process text-based symptoms, patient`s personal information and past medical history is used as input data. In the next step, data is trained for the possible detected diagnoses and recommend appropriate treatment and as a result of training we expect to get out model output.

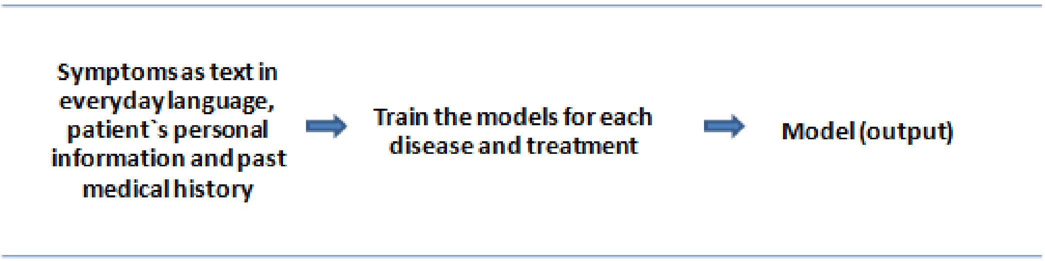
Symptoms as text in everyday language, patient`s personal information and past medical history
Train the models for each disease and treatment
Model (output)

Figure 3. Implementation steps for training process

Testing: After the training process, the system is tested for various diagnoses and patients, then the percentage risk of possible disease is represented as output information. Corresponding to this, treatments and recommendations (any medical examinations, tests) are driven.

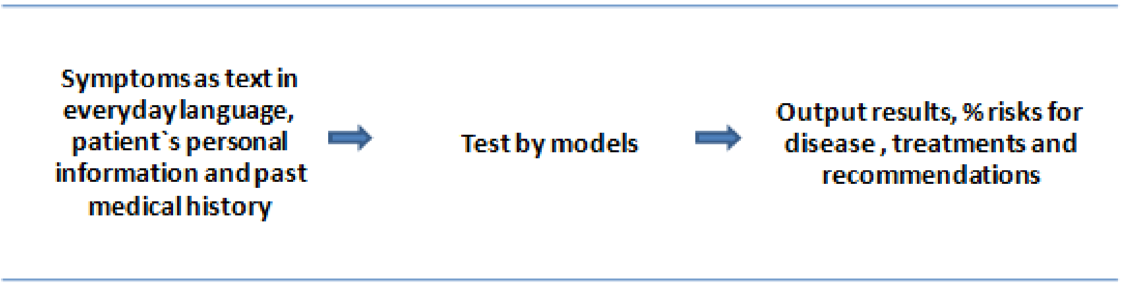
Symptoms as text in everyday language, patient`s personal information and past medical history
Test by models
Output results, % risks for disease , treatments and recommendations

Figure 4. Procedure of testing and inferences

IV. RESULTS AND CONCLUSION

Symptom analysis system, I-CURE (Intelligent context utilizing recommendation engine) is developed to increase the accuracy of medical diagnosis. The project may provide a significant help in clinical decision process which gives effective results even with patients own words as symptoms. The behavior of different classifiers was tested in the context of the problem. classifying inputted symptoms into predefined disease classes. The problem was approached with three different methods: using naïve Bayes, decision tree and Fuzzy Wuzzy library.

The goal was to make a system that will give as correct classification as possible regardless of spelling mistakes. Different inputs were tested to assess the abilities supported by the Text Blob library. Output is based on the result obtained using Fuzzy Wuzzy library regardless of some spelling mistakes that user might have done in giving input. After testing the system, an accuracy of 72.5% has been accomplished. The impact on outcomes, assessing whether the project reduces time from diagnosis to treatment, reduces cost, and improves quality the benefits of the study in global health care environment.

Future Study Objectives: Detailed information about the patient will also be collected to evaluate the model efficiently and give more accurate prediction of treatments and recommendations. These are patients past medical history (allergies, medicines, surgeries, family history, diets, and habits), birth and growth information, age, gender, height, weight. To improve the current system, we aim to upgrade I-CURE available without restriction of input language.

REFERENCES

1. Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, 17. https://doi.org/D010001275 [pii]
2. Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. *Proceedings : A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 485–9. Retrieved fromhttp://www.ncbi.nlm.nih.gov/pubmed/9357673%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2233565
3. Davis, D. a., Chawla, N. V., Blumm, N., Christakis, N., Barabasi, A.-L., & Barabási, A. (2008). Predicting individual disease risk based on medical history. *Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM '08*, 769. https://doi.org/10.1145/1458082.1458185
4. Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, *11*(5), 392–402. https://doi.org/10.1197/jamia.M1552
5. Kaul, C., Kaul, A., & Verma, S. (2015). Comparitive study on healthcare prediction systems using big data. *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 1–7. https://doi.org/10.1109/ICIIECS.2015.7193095
6. Leape, L. L., Brennan, T. A., Laird, N. A. N., Ph, D., Lawthers, A. N. N. G., Sc, D., … Hiatt, H. (1991). THE NATURE OF ADVERSE EVENTS IN HOSPITALIZED PATIENTS Results of the Harvard Medical Practice Study II fecting the quality of care has grown . Curiously , how- paratively little attention from either perspective . But an important objective for those conc, *324*(6), 377–384.
7. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *IMIA Yearbook of Medical Informatics Methods Inf Med*, *47*(1), 128–44. https://doi.org/me08010128
8. Popowich, F. (2005). Using Text Mining and Natural Language Processing for Health Care Claims Processing. *ACM SIGKDD Explorations Newsletter - Natural Language Processing and Text Mining*, *7*(1), 59–66. https://doi.org/10.1145/1089815.1089824
9. Pratt, W., & Yetisgen-Yildiz, M. (2003). A Study of Biomedical Concept Identification: MetaMap vs. People. *Journal of the American Medical Informatics Association*, 529–533. https://doi.org/D030003464 [pii]
10. Schiff, G. D. (2009). Diagnostic Error in Medicine. *Archives of Internal Medicine*, *169*(20), 1881. https://doi.org/10.1001/archinternmed.2009.333
11. Singh, H., & Sittig, D. F. (2015). Setting the record straight on measuring diagnostic errors. Reply to: "Bad assumptions on primary care diagnostic errors" by Dr Richard Young. *BMJ Quality & Safety*, *24*(5), 345.2-348. https://doi.org/10.1136/bmjqs-2015-004140
12. Spyns, P., Nhàn, N. T., Baert, E., Sager, N., & De Moor, G. (1998). Medical language processing applied to extract clinical information from dutch medical documents. *Studies in Health Technology and Informatics*. https://doi.org/10.3233/978-1-60750-896-0-685
13. Tange, H. J., Schouten, H. C., Kester, A. D., & Hasman, A. (1998). The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association : JAMIA*, *5*(6), 571–82. https://doi.org/10.1136/jamia.1998.0050571
14. Wright, L. W., Nardini, H. K. G., Aronson, A. R., & Rindflesch, T. C. (1999). Hierarchical Concept Indexing of Full-text Documents in the UMLS Information Sources Map. *Journal of the American Society for Information Science*, *50*(6), 514–523. Retrieved from http://apps.isiknowledge.com/full_record.do?product=UA&search_mode=Refine&qid=8&SID=X24eA7pAEbF36l1CgDp&page=1&doc=3&colname=WOS
15. Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, *17*(1), 19–24. https://doi.org/10.1197/jamia.M3378
16. Zou, Q., Chu, W. W., Morioka, C., Leazer, G. H., & Kangarloo, H. (2003). IndexFinder: a method of extracting key concepts from clinical texts for indexing. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 763–7. https://doi.org/D030003344 [pii]
17. Steven Loria (2017). TextBlob: Simplified text processing [a Python (2 and 3) library for processing textual data]. Retrieved from http://textblob.readthedocs.io/en/latest/index.html